Flitto Agent Bench

Finance – US Consumer Credit · 기술 보고서

Flitto Agent Data Researcher Team | 데이터 버전 2026-11-18 · 평가 기준일 2026-11-13

보고서 판 26.11.18 · 문항 세트 판본 26.11.11 · 분석·보고서 개정일 2026-11-18

초록

금융 실무에 투입되는 AI 에이전트는 여러 기록을 대조해 관련 규정을 적용하고, 정해진 예산 안에 요구된 결과물을 제출해야 합니다. Flitto Agent Bench (Finance – US Consumer Credit)는 실무 규정과 업무 흐름을 재현한 과제로 에이전트의 판단·수행 능력과 취약점을 측정합니다. 첫 공개 트랙은 미국 소비자 신용·카드 백오피스 업무를 다룹니다. 5개 모델의 개발 단계 응답을 바탕으로 8개 업무 분야의 99문항을 선별했습니다.

업무와 데이터는 가상(합성)입니다. 일부 생성기는 실제 금융 자료의 스키마와 집계 통계에 맞춰 보정했고, 판단 기준은 미국 공개 규정과 실무 자료에서 가져왔습니다. 실제 고객 기록은 들어 있지 않습니다.

핵심 지표는 필수 값과 제출 규약을 모두 지켜 예산 안에 과제를 끝낸 비율인 '예산 내 완수율'입니다. 동일한 99문항을 두 차례 평가한 평균은 Opus 79.3%, Gemini 53.5%, GPT 28.8%, Qwen 27.8%, Kimi 16.2%였습니다. 두 번 모두 통과한 비율은 10.1~75.8%, 두 회차의 판정 일치율은 78.8~92.9%였습니다. 보고서는 이 다섯 모델의 반복 평가와 Fable·GLM의 별도 비교를 담습니다. Fable·GLM도 같은 99문항을 두 차례 평가했으며, 평균 완수율은 각각 85.9%, 27.3%였습니다.

1. 연구 목적과 의의

금융 실무에서는 단순 수식 계산에 앞서 어떤 기록을 쓸지, 집계 기간과 대상 집단을 어떻게 잡을지, 어떤 정책을 적용할지부터 판단해야 합니다. 필수 항목을 빠뜨리거나 제한된 예산 안에 제출하지 못하는 것 역시 명백한 업무 실패입니다. Flitto Agent Bench는 에이전트에게 업무 기록과 규정집 묶음인 '데이터룸'을 제공하고, 명확한 제출 규약 아래에서 이러한 일련의 업무 수행 능력을 측정합니다.

본 연구의 의의는 크게 세 가지입니다. 첫째, 미국의 공개 규정과 실무 영상에서 발췌한 현업의 판단 기준을 가상 데이터룸 기반의 소비자 신용·카드 업무 과제로 구현했습니다. 둘째, 도구를 사용해 자료를 직접 탐색한 뒤 필수 결과물을 모두 제출해야 통과하는 완수 기준을 세웠습니다. 셋째, 5개 모델을 같은 99문항에서 두 차례 평가하고, 평균 완수율과 반복 성공·실패를 함께 보고합니다.

1.1 선행 연구와의 차별점

도구를 활용하는 금융 에이전트 평가는 연구마다 초점이 다릅니다. DABstep은 금융 데이터와 문서를 넘나드는 다단계 분석을 다루고, Finch는 스프레드시트 기반의 재무·회계 작업에 집중하며, τ-Knowledge의 은행 부문은 지식 검색과 도구 사용을 결합한 고객지원을 다룹니다. APEX-Agents와 GDPval은 금융을 포함한 전문직 산출물의 품질을 봅니다. 이들과 비교했을 때 Flitto Agent Bench는 도메인뿐만 아니라 최종 제출물과 채점 기준에서 뚜렷한 차이를 보입니다.

표 1. 관련 평가의 과제와 채점 기준

연구	주요 과제	제출 형태 및 채점 기준
DABstep	금융 데이터·문서의 다단계 분석	수치·문자·목록형 답의 자동 비교
Finch	스프레드시트 중심 재무·회계 실무	파일 생성·수정 및 질의응답, 전문가·자동 채점 병행
τ-Knowledge	은행 고객지원을 위한 검색 및 도구 활용	대화 진행과 도구 실행에 따른 목표 상태 도달 여부
APEX-Agents	투자은행·컨설팅·법률 분야 다중 앱 활용	응답 및 생성 파일의 필수 요건 충족 여부
GDPval	다양한 직군의 실제 지식 노동	전문가의 블라인드 산출물 품질 비교
Flitto Agent Bench	미국 소비자 신용·카드 업무의 기록 대조·정책 적용	필수 값 정확도, 제출 규약 준수, 실행 예산 내 완료 동시 충족

기존 연구의 금융 분석, 문서 작성, 대화형 상태 변경과 비교하면, 본 벤치마크는 고정된 원천 자료에서 요청 범위를 판단해 규약에 맞는 최종 답안을 예산 안에 제출하는 과정에 초점을 둡니다.

2. 데이터 및 문항 설계

2.1 가상 데이터 구축과 실무 암묵지 반영

업무와 데이터는 가상(합성)입니다. 일부 생성기는 실제 금융 자료의 스키마와 집계 통계에 맞춰 보정했고, 판단 기준은 미국 공개 규정과 실무 자료에서 가져왔습니다. 실제 고객 기록은 들어 있지 않습니다.

특히 문항을 설계할 때 실무자 인터뷰, 의회 청문회, 웨비나, 교육 영상 등에 녹아 있는 현장 노하우(암묵지)를 활용했습니다. 여기서 암묵지란 실무자의 설명 속에 자연스럽게 드러나는 자료 해석 방식, 판단 순서, 예외 상황 처리 등을 말합니다. 활용한 자료에는 신용 심사 전략을 다룬 전문가 대담과 신용 보고 정확성에 관한 미 하원 공개 청문회가 포함되며, 이를 과제의 배경과 세부 요구사항으로 구체화했습니다.

개별 과제는 원천 업무 기록, 정책·절차 문서, 업무 지시서, 출력 규약, 채점 기준, 실행 예산으로 구성됩니다. 신용 심사, 신용파일 관리, 결제·청구 등 총 8개 실무 분야를 아우르며, 문항마다 핵심 분야를 하나씩 배정했습니다.

표 2. 업무 분야별 문항 구성

업무 분야	문항	기본 설계
개인 금융	13	8
결제·청구	13	10
사기·분쟁	12	9
신용파일	11	7
심사	13	11
이탈·유지	13	9
추심	12	8
포트폴리오	12	12
합계	99	74

최종 수록된 99개 문항은 74개의 기본 설계(과제의 공통 뼈대)에서 파생되었습니다. 입력 데이터는 63개의 데이터룸 패키지로 제공되며, 일부 문항은 기본 설계나 데이터룸을 공유합니다.

2.2 개발 응답을 활용한 문항 선별

후보 311문항에서 Opus·Gemini·GPT·Qwen·Kimi의 개발 단계 응답을 바탕으로 99문항을 골랐습니다. 업무 분야별 구성을 함께 고려했습니다. 개발 단계의 통과 수는 부록 B에 제시합니다.

모델이 어려워하는 문항을 선별하는 접근은 HLE와 ANLI에도 쓰입니다. 본 벤치마크는 실제 업무에서 필요한 자료 해석, 정책 적용, 계산과 제출을 평가 대상으로 삼습니다.

2.3 정답 검증 및 사후 점검

문항 제작 단계에서 금융 전문가가 업무 논리를 검토해 정답 판정에 동의했고, 팩트체크도 거쳤습니다. 외부 규칙의 근거는 미국의 공개 1차 출처 원문과 대조했고, 데이터룸 문서를 인용한 내용이 제공된 문서에 실제로 있는지도 확인했습니다. 정답 도출을 위한 계산식과 출처 대조 내역은 별도 근거 자료로 관리합니다.

정답은 지시서의 요구 범위, 데이터 단위와 기간, 규정 및 기록 간 우선순위에 근거합니다. 98개 문항은 하나의 정답을, 1개 문항은 정의상 허용되는 두 가지 답을 가지며, 정답 형태는 수치와 문자열을 모두 포함합니다.

모델명을 가린 독립 검토자가 최종 99개 문항의 미분류 실패 사례를 대상으로 대안 해석 가능성을 점검했습니다. 저장된 실행 기록과 채점 결과도 독립적으로 대조해 점수 집계와 통계 처리를 재현했습니다. 정답은 문항별 풀이·정책 근거와 설계 출처에 연결해 관리합니다.

3. 평가 방법

3.1 평가 대상과 공통 조건

본 평가 대상은 Claude Opus 5.5(Opus), Gemini 4 Argon(Gemini), GPT-6 Astra(GPT), Qwen3.8 Max 0902(Qwen), Kimi K3.1 Pro(Kimi)입니다.

모든 모델에 동일한 99문항, 데이터룸, 출력 규약과 채점 기준을 적용했습니다. 문항당 1회 시도하며, 예산은 최대 60턴·3,600초입니다. 일곱 모델을 각각 두 차례 평가하고 주 점수는 두 회차의 평균으로 계산했습니다.

입력 조건이 일치하는 첫 번째 유효 실행을 채점했습니다. 오답·미제출·예산 초과는 실패입니다. 사전에 정한 인프라 오류에는 1회 재실행을 허용했습니다. 문항·프롬프트·데이터룸 파일의 해시와 채점기 버전이 고정 세트와 일치하는 실행만 채점했습니다.

평가 수행일 (시작일–종료일)

모델	1회차	2회차
Opus	2026-11-08–2026-11-11	2026-11-09–2026-11-11
Gemini	2026-11-05–2026-11-11	2026-11-09–2026-11-11
GPT	2026-11-08–2026-11-11	2026-11-09–2026-11-11
Qwen	2026-11-08–2026-11-11	2026-11-09–2026-11-11
Kimi	2026-11-08–2026-11-11	2026-11-09–2026-11-11
Fable	2026-11-11	2026-11-13
GLM	2026-11-08–2026-11-11	2026-11-13

3.2 채점 기준 및 핵심 지표

에이전트는 요구된 모든 필수 키(Key)를 담아 답안 파일을 제출해야 합니다. 코드 베리파이어는 파일 존재 여부, 값의 정확도, 프로세스 정상 종료 여부를 각각 검사합니다. LLM 심판은 사용하지 않습니다. 모든 항목이 규약에 맞고 예산 내에 정상 종료되었을 때만 '완수'로 인정합니다. 답안 내용이 정확하더라도 턴이나 시간 한도에 걸려 강제 종료되면 실패로 간주합니다. 일부 필드가 맞더라도 문항 단위의 부분 점수는 없으며 미제출, 필수 키 누락, 포맷 오류 역시 모두 오답 처리합니다.

수치 항목은 숫자형 데이터나 값이 같은 숫자 문자열만 허용합니다. 건수에는 정수값만 허용하고 금액이나 비율에는 문항별 허용 오차를 두었으며, 쉼표(,)나 통화 기호($)가 섞인 문자열, 불리언(Boolean) 값 등 규약 밖의 값은 인정하지 않습니다. 문자열 항목은 문항별 규약을 따릅니다. 각 문항의 필수 항목을 하나의 정답 기준에 맞춰 모두 충족해야 하며, 같은 필드에 서로 다른 값을 함께 기재하면 오답 처리합니다.

- 주 완수율: 두 회차의 문항당 1회 시도 완수율을 평균한 값입니다. (1회차 통과 수 + 2회차 통과 수) / (2 × 99) × 100으로 계산합니다.
- 반복 완수율 (pass^2): 같은 문항을 두 번 모두 통과한 비율입니다.
- 2회차 일치율: 두 번 모두 통과했거나 두 번 모두 실패해 1·2회차 판정이 그대로 유지된 비율입니다.

3.3 통계 분석

부록 A의 1회차 완수율에 대한 95% 신뢰구간은 문항 단위 10,000회 복원추출을 거친 백분위 부트스트랩(seed 20261104)으로 구했습니다. 이 구간의 대상은 선정된 99문항의 1회차 결과입니다. 부록 C의 업무 분야별 1회차 구간은 Wilson 방식(z=1.96, 연속성 보정 미적용)으로 계산했습니다.

모델 비교에는 같은 문항의 통과·실패 쌍에 대한 정확 맥니머 검정을 사용하고, 회차별 다섯 모델의 10쌍에 Holm 보정(유의수준 0.05)을 적용했습니다. 등급은 각 회차의 점수 순서에서 인접 모델 간 유의한 차이로 경계를 구한 뒤, 두 회차에 공통으로 나타난 경계를 유지했습니다. 회차별 이웃 쌍과 공통 경계는 부록 A에 제시합니다.

4. 평가 결과

4.1 두 회차 평균 완수율

표 3. 동일한 99문항의 두 회차 평균과 등급

모델	1회차 통과/99	2회차 통과/99	평균 완수율 %	등급
Claude Opus 5.5	81	76	79.3	1
Gemini 4 Argon	50	56	53.5	2
GPT-6 Astra	29	28	28.8	3
Qwen3.8 Max 0902	26	29	27.8	3
Kimi K3.1 Pro	14	18	16.2	3

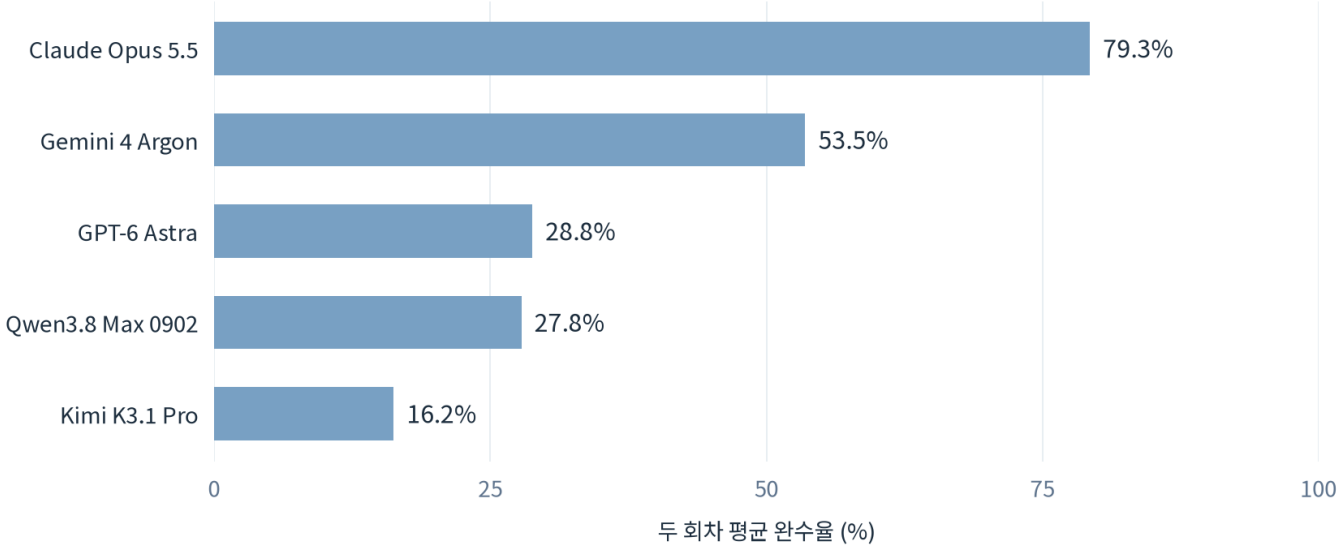

그림 1. 두 회차 평균 완수율. 막대와 수치는 같은 99문항의 두 회차 평균입니다.

1회차에서는 다섯 모델이 모두 실패한 문항이 12개, 모두 통과한 문항이 7개였습니다. 나머지 80문항에서 모델 간 결과가 갈렸습니다. 1~3개 모델만 통과한 문항은 74개였습니다.

1회차의 예산 초과 종료는 Gemini 19개와 Qwen 2개였습니다. Gemini의 19개 중 18개는 답안 파일을 제출하지 못했고, 1개는 답안을 남겼으나 예산 내 정상 종료 조건을 충족하지 못했습니다.

4.2 반복 평가 일치율과 등급

동일 문항을 2회 평가한 결과는 다음과 같습니다.

표 4. 반복 완수율과 판정 일치율

모델	1회차 %	2회차 %	평균 %	pass^2 %	일치율 %
Opus	81.8	76.8	79.3	75.8	92.9
Gemini	50.5	56.6	53.5	44.4	81.8
GPT	29.3	28.3	28.8	19.2	80.8
Qwen	26.3	29.3	27.8	17.2	78.8
Kimi	14.1	18.2	16.2	10.1	87.9

1회차에는 GPT가 Qwen보다 높은 점수를 기록했고, 2회차에는 Qwen이 GPT보다 높았습니다. Qwen은 1회차 26개, 2회차 29개 문항을 통과했으며, 두 번 모두 통과한 문항은 17개(17.2%)였습니다. 통과에서 실패로 바뀐 문항은 9개, 실패에서 통과로 바뀐 문항은 12개였습니다. Kimi의 일치율(87.9%)에는 두 회차 모두 실패한 문항도 포함됩니다. 회차 간 정오 변화를 비교한 정확 맥니머 검정 p값은 모델별로 0.125~1.000이었습니다.

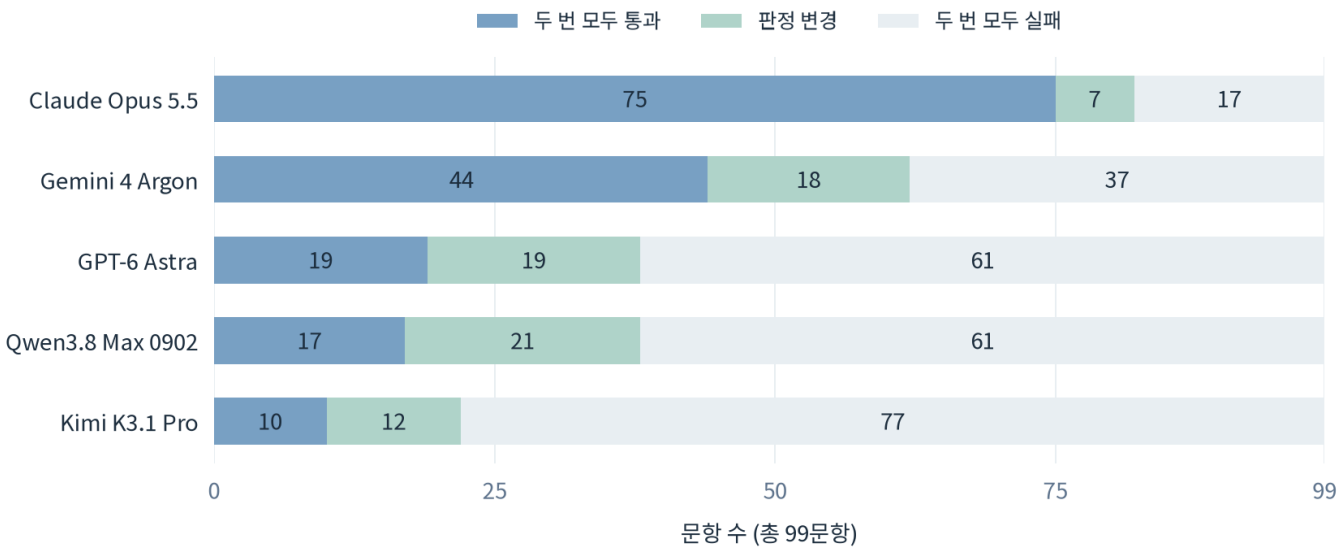

그림 2. 같은 99문항의 반복 판정. 두 번 모두 통과, 판정 변경, 두 번 모두 실패한 문항 수입니다.

두 회차 공통 등급은 {Opus} > {Gemini} > {GPT, Qwen, Kimi}입니다. GPT와 Qwen의 1회차 차이는 유의하지 않았습니다(Holm p=0.728). 1회차에서 유의했던 GPT–Kimi와 Qwen–Kimi의 차이는 2회차에서 유의하지 않아 GPT·Qwen·Kimi가 같은 3등급으로 묶였습니다.

4.3 실행 기록 검수

저장된 실행 기록과 채점 결과를 대조해 집계와 통계 처리를 재현했습니다. 실행 기록 감사에서는 정답·채점기·벤치마크 저장소·다른 실행 자료에 대한 접근이 관측되지 않았습니다. 두 번째 평가의 495개 실행에서도 같은 항목을 점검했습니다.

모델명과 통과 여부를 가린 분류자가 실행 기록의 평가 인식 표현을 같은 기준으로 분류했습니다. 해당 표현은 1회차 Opus 9개, Gemini 1개, Qwen 1개와 2회차 Opus 3개, Qwen 1개 실행에서 관측됐습니다. 그중 통과한 실행을 실패로 놓는 민감도 분석에서 1회차 통과 수는 Opus 77개·Gemini 49개, 2회차는 Opus 75개·Qwen 28개였습니다. 이 분석에서 2회차 Qwen과 GPT는 28개로 동점이었습니다.

4.4 문항 선정에 사용하지 않은 모델의 별도 비교

Claude Fable 5.2와 GLM-5.3 Prime을 같은 99문항, 출력 규약, 60턴·3,600초 예산과 채점 기준으로 각각 두 차례 평가했습니다.

표 5. 별도 비교 모델의 두 회차 결과

모델	1회차 통과/99	2회차 통과/99	평균 완수율 %	최저–최고 %
Claude Fable 5.2	85	85	85.9	85.9–85.9
GLM-5.3 Prime	29	25	27.3	25.3–29.3

Fable의 1회차에서는 98개 실행이 정상 종료했고 1개가 60턴 한도로 종료됐습니다. 1회차 문항 부트스트랩 95% 신뢰구간은 78.8~91.9%였습니다. 본 평가 모델들의 1회차와 같은 문항끼리 비교해, Fable만 통과한 문항은 Opus 대비 7개, Opus만 통과한 문항은 3개였습니다(p=0.344). Gemini·GPT·Qwen·Kimi와의 차이는 각각 Holm p<0.001이었습니다.

GLM의 1회차는 Opus·Gemini보다 낮고 Kimi보다 높은 완수율을 보였습니다(Holm p 각각 <0.001, 0.002, 0.008). GPT·Qwen과의 차이는 유의하지 않았습니다(각각 Holm p=1.000). 이 비교의 Holm 보정은 Fable과 GLM 각각 본 평가 다섯 모델과의 5쌍에 적용했습니다.

4.5 실패 유형: 제출 답안과 실행 상태

분석 범위는 일곱 모델이 동일한 99문항을 두 차례 수행한 1,386건입니다. 이 중 631건이 통과하고 755건이 실패했습니다. 다섯 본 평가 모델의 실패 583건과 Fable·GLM 별도 비교의 실패 172건을 진단 목적으로 합산했습니다. 같은 문항이라도 모델과 회차가 다르면 별도 실행으로 셉니다. 두 번 중 한 번 이상 통과한 문항을 세는 공개 리더보드의 pass@2와는 분모가 다릅니다.

실패 755건 중 392건(51.9%)의 답안에서 자료 해석·대조 관련 오답 패턴이 확인됐습니다. 데이터 이상을 반영하지 않거나 다른 근거보다 원본 필드 값을 우선한 사전 정의 오답 패턴과 일치한 실행입니다. 대상·잔액 정의 관련 패턴은 158건, 답할 수 있는 값의 보류는 65건에서 나타났습니다.

실패 유형의 구성은 모델마다 다릅니다. Claude Opus 5.5에서는 대상·잔액 정의 패턴이 실패 41건 중 24건(58.5%), Claude Fable 5.2에서는 실패 28건 중 22건(78.6%)으로 가장 많았습니다. 전체 평가 실행 기준 발생률은 각각 24/198(12.1%)과 22/198(11.1%)입니다. 웹의 모델별 실패 진단에서 실패 실행 기준과 전체 평가 실행 기준을 전환할 수 있습니다.

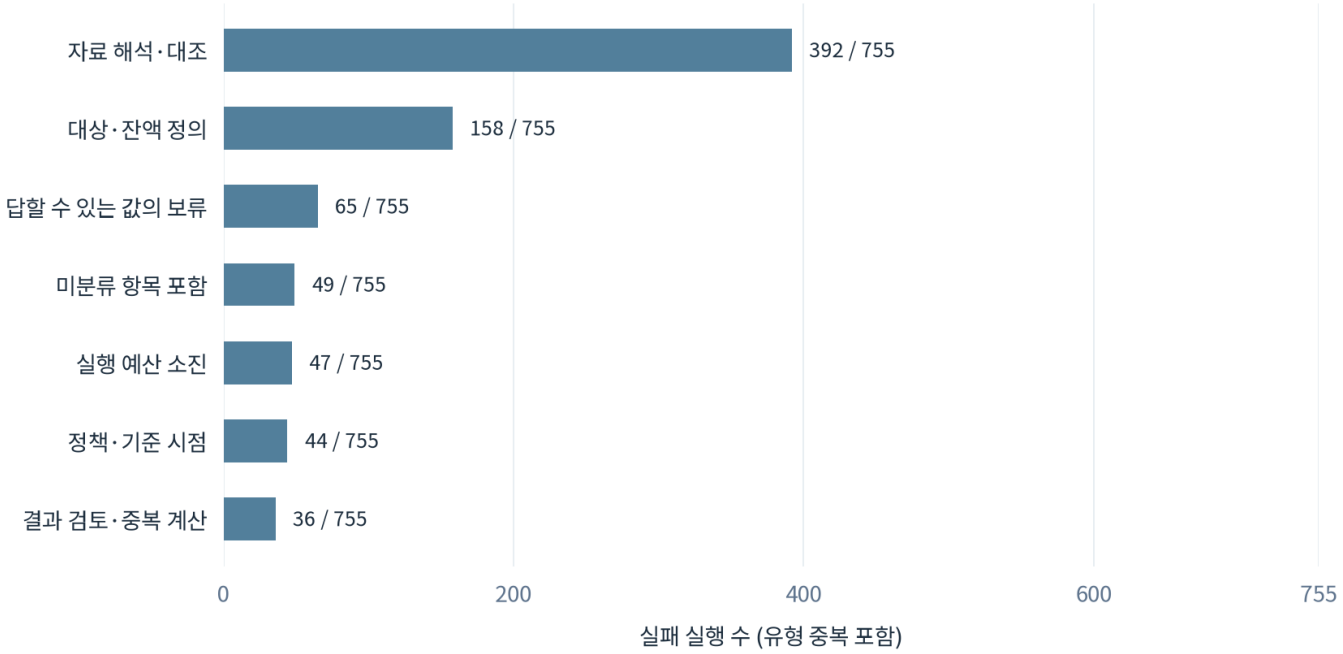

그림 3. 두 회차 실패 실행의 진단 유형. 실패 755건 기준이며 한 실행이 여러 유형에 포함될 수 있습니다.

표 6. 실패 유형별 실행 수와 평가 집단 (중복 포함)

진단 유형	본 평가 5개 모델	별도 비교 2개 모델	전체
자료 해석·대조	304	88	392
대상·잔액 정의	109	49	158
답할 수 있는 값의 보류	61	4	65
미분류 항목 포함	36	13	49
실행 예산 소진	41	6	47
정책·기준 시점	32	12	44
결과 검토·중복 계산	24	12	36
실패 실행 수 (분모)	583	172	755

자료 해석·대조는 데이터 이상과 필드 간 우선순위, 대상·잔액 정의는 집계 대상과 잔액의 정의, 정책·기준 시점은 적용 규칙과 보고 시점에 관한 오답 패턴을 묶었습니다. 결과 검토·중복 계산은 초안 검토·대사와 중복 집계 관련 패턴입니다. 답할 수 있는 값의 보류는 수치로 답할 수 있는 필수 항목에 보류 값을 제출한 경우입니다. 유형마다 실행 한 건을 한 번만 세지만, 같은 실행이 여러 유형에 포함될 수 있으므로 행의 합은 분모보다 큽니다.

실패 중 708건은 답안을 제출하고 정상 종료했지만 필수 검사를 통과하지 못했습니다. 나머지 47건은 실행 예산을 소진했습니다. 이 47건 중 42건은 답안 파일이 없었고, 5건은 파일을 남겼습니다. 파일을 남긴 5건 중 3건은 답안 검사를 모두 통과했어도 예산 내 정상 종료 조건을 충족하지 못해 실패했습니다. 미분류 항목이 있는 49건 중 46건은 전체가 미분류이며, 3건은 확인된 패턴과 미분류 항목이 함께 있습니다.

이 분석은 저장된 답안과 종료 상태를 고정된 채점 규칙 및 등록된 오답 패턴에 대조했습니다. 1,386건의 통과·실패 판정은 이 보고서의 집계와 모두 일치하며, 이 진단은 점수를 바꾸지 않습니다. 진단은 답안의 패턴 일치에 근거하므로 모델이 실제로 어떤 자료를 읽거나 건너뛰었는지, 어떤 순서로 판단했는지를 확정하지 않습니다. 개별 행동 원인을 확인하려면 실행 기록을 별도로 검토해야 합니다. 개별 문항, 오답 패턴의 구체적 조건, 정답과 실행 기록은 공개하지 않습니다.

5. 논의

Flitto Agent Bench는 미국 소비자 신용·카드 업무의 판단과 산출물 제출 과정을 가상 데이터룸 위에 구현했습니다. 일부 생성기는 실제 금융 자료의 스키마와 집계 통계로 보정했고, 공개 규정과 실무자 영상에서 선별한 현장 지식을 문제 조건에 반영했습니다. 이를 통해 에이전트가 정해진 예산 안에 필수 결과물을 제출하는지 평가합니다.

실패 진단은 제출 답안과 실행 기록을 대조해 사람이 검토할 후보를 제공합니다. 재현 기록에는 채택 문항, 입력·설정 식별 정보, 실행의 종료 상태, 채점 판정, 집계·통계 절차를 연결했습니다. 비공개 평가 문항과 정답, 고객 자료의 공개 범위는 집계 결과의 보고 범위와 분리해 관리합니다.

6. 결론

Flitto Agent Bench는 업무 자료의 해석부터 정책 적용, 계산, 규약에 맞춘 제출까지 전체 과정을 평가합니다. 같은 99문항을 두 차례 평가한 일곱 모델의 평균 완수율은 16.2~85.9%였습니다. 반복 판정과 모델 간 비교는 이 과제 집합에서 관측된 수행 차이를 보여 줍니다.

플리토는 이와 별도로 AI 학습 데이터를 공급합니다.

7. 공개 화면의 자원 지표

공개 화면은 도구 호출, 입력 토큰, 출력 토큰, 실행 시간을 제공합니다. 사용량 막대는 문항별 1회차 기록을 집계합니다. 평균은 기록이 있는 문항 수로 나누고, 합계는 모든 문항의 기록이 있을 때 표시합니다. 파레토 차트는 선택한 집계 방식에서 같은 99문항의 점수와 사용량 기록이 모두 있는 모델을 비교합니다. 점선은 자원을 더 쓰지 않고 점수를 높이거나, 점수를 낮추지 않고 자원을 줄이는 다른 선택 모델이 없는 지점을 잇습니다.

부록 A. 신뢰구간 및 모델 간 일대일 비교

표 A1. 1회차 완수율과 문항 부트스트랩 95% 신뢰구간 (%)

문항 복원추출 10,000회, seed 20261104 기준 백분위 구간입니다.

모델	1회차 완수율 %	95% 구간
Opus	81.8	73.7 - 88.9
Gemini	50.5	40.4 - 60.6
GPT	29.3	20.2 - 38.4
Qwen	26.3	18.2 - 35.4
Kimi	14.1	8.1 - 21.2

표 A2. 1회차 정확 맥니머 검정 및 Holm 보정 결과

(가운데 두 열은 비교 쌍 중 해당 모델만 단독으로 통과한 문항 수)

모델 쌍	앞 모델만	뒤 모델만	정확 p	Holm p
Opus-Gemini	33	2	<0.001	<0.001
Opus-GPT	54	2	<0.001	<0.001
Opus-Qwen	59	4	<0.001	<0.001
Opus-Kimi	67	0	<0.001	<0.001
Gemini-GPT	26	5	<0.001	<0.001
Gemini-Qwen	30	6	<0.001	<0.001
Gemini-Kimi	39	3	<0.001	<0.001
GPT-Qwen	18	15	0.728	0.728
GPT-Kimi	19	4	0.003	0.008
Qwen-Kimi	17	5	0.017	0.034

표 A3. 회차별 이웃 쌍 검정

(Holm 보정 범위는 회차별 다섯 모델의 열 쌍입니다. 각 회차의 점수 순서에서 이웃한 쌍과 두 회차 공통 등급 경계를 각각 제시합니다.)

회차	이웃 쌍	Holm p	해당 회차 유의
1	Opus-Gemini	<0.001	예
1	Gemini-GPT	<0.001	예

회차	이웃 쌍	Holm p	해당 회차 유의
1	GPT-Qwen	0.728	아니요
1	Qwen-Kimi	0.034	예
2	Opus-Gemini	0.002	예
2	Gemini-Qwen	<0.001	예
2	Qwen-GPT	1.000	아니요
2	GPT-Kimi	0.174	아니요

표 A4. 두 회차 공통 등급 경계

공통 등급 경계	1회차 이웃 쌍	2회차 이웃 쌍
{Opus} / {Gemini, GPT, Qwen, Kimi}	Opus-Gemini	Opus-Gemini
{Opus, Gemini} / {GPT, Qwen, Kimi}	Gemini-GPT	Gemini-Qwen

부록 B. 개발 단계 성적

표 B1. 후보 문항 선별 여부별 개발 단계 통과 수 / 채점 수

개발 통과율 순서는 세 집합 모두 Gemini, Opus, Qwen, GPT, Kimi였습니다.

문항 집합	Opus	Gemini	GPT	Qwen	Kimi
선정 99	43/99	50/99	18/99	35/99	10/99
제외 212	150/212	156/212	111/212	140/212	101/212
개발 전체 311	193/311	206/311	129/311	175/311	111/311

부록 C. 1회차 통과 분포와 업무 분야별 성적

표 C1. 1회차에서 통과한 모델 수에 따른 문항 분포

통과 모델 수	문항	Opus	Gemini	GPT	Qwen	Kimi
0	12	0	0	0	0	0
1	30	25	1	1	3	0
2-3	44	43	36	16	12	4
4-5	13	13	13	12	11	10

표 C2. 1회차 업무 분야별 통과 수 / 문항 수 [Wilson 95% 신뢰구간, %]

업무 분야	Opus	Gemini	GPT	Qwen	Kimi
개인 금융	5/13 [17.7, 64.5]	7/13 [29.1, 76.8]	4/13 [12.7, 57.6]	6/13 [23.2, 70.9]	1/13 [1.4, 33.3]
결제·청구	10/13 [49.7, 91.8]	7/13 [29.1, 76.8]	4/13 [12.7, 57.6]	4/13 [12.7, 57.6]	3/13 [8.2, 50.3]
사기·분쟁	11/12 [64.6, 98.5]	6/12 [25.4, 74.6]	2/12 [4.7, 44.8]	1/12 [1.5, 35.4]	1/12 [1.5, 35.4]
신용파일	10/11 [62.3, 98.4]	6/11 [28.0, 78.7]	2/11 [5.1, 47.7]	5/11 [21.3, 72.0]	3/11 [9.7, 56.6]
심사	13/13 [77.2, 100.0]	9/13 [42.4, 87.3]	6/13 [23.2, 70.9]	1/13 [1.4, 33.3]	0/13 [0.0, 22.8]
이탈·유지	13/13 [77.2, 100.0]	5/13 [17.7, 64.5]	6/13 [23.2, 70.9]	4/13 [12.7, 57.6]	2/13 [4.3, 42.2]
추심	10/12 [55.2, 95.3]	5/12 [19.3, 68.0]	1/12 [1.5, 35.4]	3/12 [8.9, 53.2]	0/12 [0.0, 24.3]
포트폴리오	9/12 [46.8, 91.1]	5/12 [19.3, 68.0]	4/12 [13.8, 60.9]	2/12 [4.7, 44.8]	4/12 [13.8, 60.9]

참고문헌

- DABstep: Data Agent Benchmark for Multi-step Reasoning.
- Finch: Benchmarking Finance & Accounting across Spreadsheet-Centric Enterprise Workflows.
- τ-Knowledge: Evaluating Conversational Agents over Unstructured Knowledge.
- APEX-Agents.
- GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks.
- Humanity's Last Exam.
- Adversarial NLI: A New Benchmark for Natural Language Understanding.
- GDS Link: Credit Risk Expert Discusses the Art of Underwriting Strategy.
- U.S. House Committee on Financial Services: Consumer Credit Reporting: Assessing Accuracy and Compliance.

문항 세트 지문 (SHA-256): 72b60695c53f6e395464f2b37c383430d7aec7a770f9267b5f245bac5c469636